



RAG language models and pretrained transformer-based language models

Source: Conversation with Bing Copilot, 10.02.2025

You

What is the basic difference between a language *model based on Retrieval Augmented Generation (RAG)* and a language *model based on a Generative Pretrained Transformer (GPT)*?

Copilot

Great question! Here's a breakdown of the basic differences between a RAG (Retrieval-Augmented Generation) language model and a pretrained transformer-based language model:

Pretrained Transformer-Based Language Model

- **What it is:** It's a model that has been trained on vast amounts of text data using a transformer architecture.
- **How it works:** These models learn patterns and structures in the language during the pretraining phase and can generate text or understand language by leveraging this learned knowledge.
- **Strengths:** They excel in generating coherent and contextually relevant text, understanding context, and performing a wide range of NLP tasks.
- **Limitations:** They rely solely on the information encoded in their parameters during pretraining and do not update their knowledge base dynamically.

Retrieval-Augmented Generation (RAG) Language Model

- **What it is:** RAG is a hybrid model that combines pretrained language models with a retrieval mechanism.
- **How it works:** It retrieves relevant information from external sources (like a database or the internet) and uses this retrieved information to generate more accurate and contextually relevant responses.
- **Strengths:** They can access up-to-date information and provide more accurate and fact-based responses, especially for queries that require current or specialized knowledge.
- **Limitations:** They rely on the quality and relevance of the retrieved information and might have higher computational requirements due to the retrieval step.

In summary, while pretrained transformer-based models rely solely on their internal knowledge acquired during pretraining, RAG models enhance their capabilities by integrating external information retrieval. This allows RAG models to provide more accurate and current responses.